Master studije
GEOINFORMATIKA

Mašinsko učenje - linearna regresija

Tatjana Jakšić Krüger

tatjana@turing.mi.sanu.ac.rs

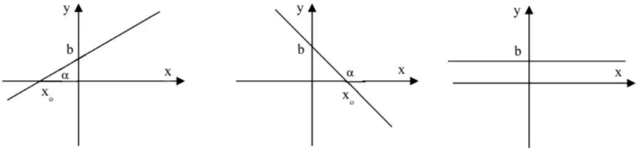
9. novembar 2021.

Građevinski fakultet Univerziteta u Beogradu

- Linearna funkcija ima oblik:

$$f(x) = ax + b, \text{ gde je } a, b \in \mathcal{R} \text{ i } a \neq 0.$$

  - *a* je *nagib* ili *kodeficijent pravca*, a *b* je *odsečak* na *y*-osi.



a > 0
права гради оштар угао са позитивним делом x-осе

a < 0
права гради туп угао са позитивним делом x-осе

a = 0
права је паралелна са x-осом

Izvor `https://www.slideshare.net/Violeta_Ivkovic/linearna-funkcija`

Osnovna terminologija, pojmovi

- U matematici promenljiva se označava sa x, y ili $\vec{x}$ da bi označili vektor.
- U statistici promenljiva je slučajna i označava se sa X, Y.
  - Raspodela, distribucija

$$F_X(x) = P(X \leq x)$$

- U mašinskog učenju promenljivu nije lako formalizovati. Moramo da znamo problem koji rešavamo.
- Veliki slovo notacija. X može da znači jednu vrednost ili vektor. Često $\boldsymbol{X}$ označava matricu.
- Posmatranja koja su označen (labeled) označavamo sa (x_i, y_i).

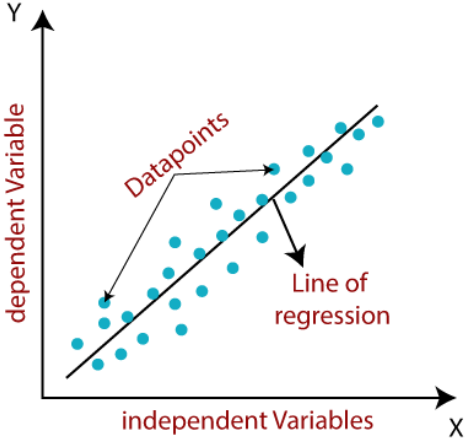
Regresiona analiza

- Skup statističkih procedura pomoću kojih ocenjujemo povezanost između nezavisne (X) i zavisne promenljive (Y).
- X je prediktorska promenljiva, faktor, regresor. Y je kriterijumska ili zavisna promenljiva.
- Rezultati pokazuju za koliko se vrednost promenljive Y menja kada se vrednost X promeni za jedinicu.
- Podela metoda regresione analize prema prirodi zavisnosti između X i Y:
  - Linearna,
  - Nelinearna regresija:
    - Kvadranta,
    - Polinomna,
    - Eksponencijalna regresija, itd.

Regresiona analiza - podela

- Prema broju nezavisnih promenljivih u regresionom modelu postoji:
  - Prosta regresija.
  - Višestruka regresija.
- Podela regresione prema vrsti zavisne promenljive:
  - Regresioni model sa neprekidnom zavisnom promenljivom.
  - Regresioni model sa kategoričkom zavisnom promenljivom.
  - Regresioni model sa dihotomnom zavisnom promenljivom.

Primer za linearni model



Linearni model

- Koristićemo slučajne promenljive X i Y. Vrednosti a i b su skalari.

$$Y = aX + b$$

- Ovaj model podrazumeva da je Y potpuno određeno sa vrednostima X.
- Pristup koji je radi:

$$E[Y|X] = aX + b$$

  gde je $E[Y|X]$ matematičko očekivanje .
- Konačno, jer $E[E[Y|X]] = E[Y]$, tada važi:

$$E[Y] = aE[X] + b.$$

- Sada znamo da želimo estimatore za parametre linearnog modela.

Estimatori linearnog modela

- Neka je θ parametar i $\hat{\theta}$ njegov estimator.
- Osobine estimatora:
  - Nepristrasan (unbiased): $E[\hat{\theta}] = \theta$
  - Minimalna varijabilnost: pronaći najmanje $var[\hat{\theta}]$.
  - Robustnost: par "odlutalih" podataka (outlier) ne bi trebalo da utiče na naše procene.
- Poznati estimamatori su: srednja vrednost (mean), medijana (median), *trimmed mean*.
- Još poznatiji estimator je *ocenjivanje metodom maksimalne verodostojnosti*. (eng. maximum likelihood estimation)

Metoda maksimalne verodostojnosti

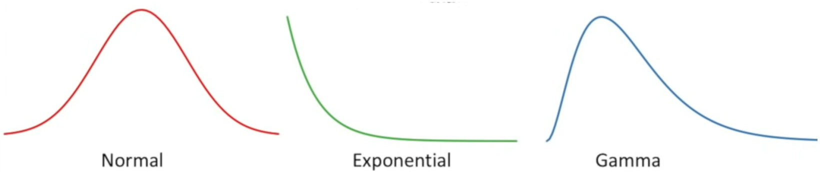
- Neka je dat skup podataka $x_1, x_2, ..., x_n$.
- Metoda max. verodostojnosti pokušava da odgovori na sledeće pitanje: koja vrednost parametra θ maksimizuje verovatnoću generisanja identičnog skupa podataka?
- Moramo da pretpostavimo da su merenja nezavisna. Neka su definisane verovatnoće $P(X_1 = x_1|\theta), P(X_2 = x_2|\theta), ..., P(X_n = x_n|\theta)$.
- U tom slučaju definišemo funkciju verodostojnosti (likelihood function) kao proizvod ovih verovatnoća.
- Da bismo evaluirali funkciju moramo da pretpostavimo raspodelu za slučajne promenljive $X_1, X_2, ..., X_n$.

Primer - Metoda maksimalne verodostojnosti

- Neka je skup podataka $x_1, x_2, ..., x_n$ merena težina kod miševa.
- Cilj - pronađi optimalan način da se fituje raspodela ovih merenja.
- Raspodela podataka omogućuje primenu iste procedure za više eksperimenata istog tipa.

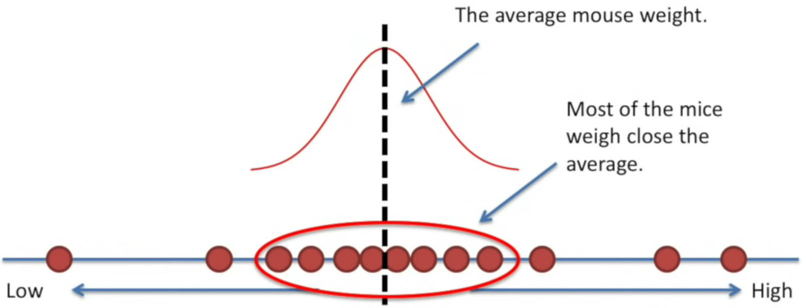
Primer - Metoda maksimalne verodostojnosti

- Neka je skup podataka $x_1, x_2, \ldots, x_n$ merena težina kod miševa.
- Cilj - pronađi optimalan način da se fituje raspodela ovih merenja.



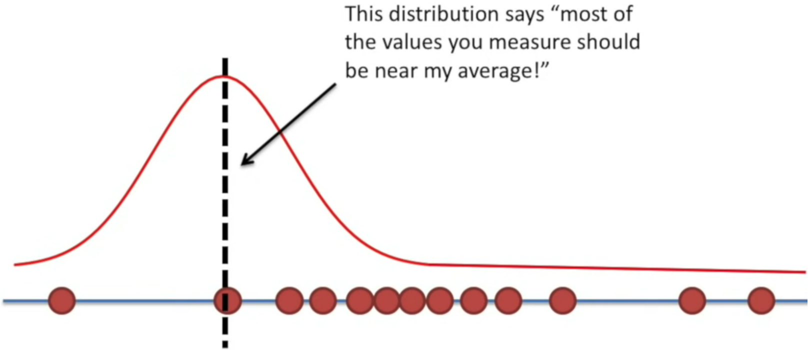
Primer - Metoda maksimalne verodostojnosti

- Neka je skup podataka $x_1, x_2, ..., x_n$ merena težina kod miševa.
- Cilj - pronađi optimalan način da se fituje raspodela ovih merenja.



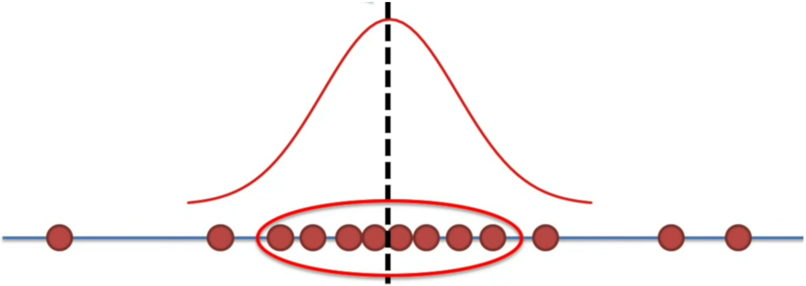
Primer - Metoda maksimalne verodostojnosti

- Neka je skup podataka $x_1, x_2, ..., x_n$ merena težina kod miševa.
- Cilj - pronađi optimalan način da se fituje raspodela ovih merenja.

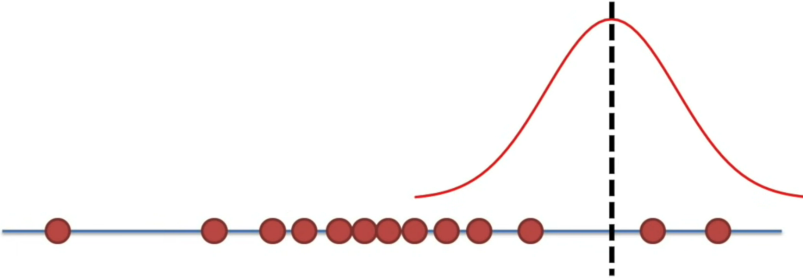


Primer - Metoda maksimalne verodostojnosti

- Neka je skup podataka $x_1, x_2, ..., x_n$ merena težina kod miševa.
- Cilj - pronađi optimalan način da se fituje raspodela ovih merenja.

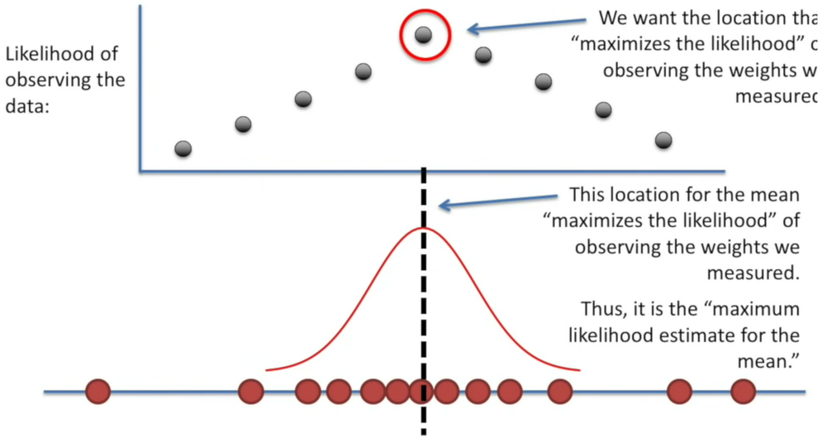


- Neka je skup podataka $x_1, x_2, ..., x_n$ merena težina kod miševa.
- Cilj - pronađi optimalan način da se fituje raspodela ovih merenja.



Primer - Metoda maksimalne verodostojnosti

- Neka je skup podataka $x_1, x_2, \dots, x_n$ merena težina kod miševa.
- Cilj - pronađi optimalan način da se fituje raspodela ovih merenja.



Metoda maksimalne verodostojnosti

- Pretpostavimo Gausovsku raspodelu $X_i \sim N(\mu, \sigma^2)$ za sve podatke i nezavisnost (*eng.* i.i.d.- independent and identically distributed).
- Za svako X_i verovatnoća da će slučajna promenljiva uzeti vrednost x_i ukoliko je vrednost parametra μ jednako je:

$$P(X_i = x_i|\mu) = \frac{1}{\sqrt{2\pi\sigma}} \exp^{-(x_i-\mu)^2/2\sigma^2} . \tag{1}$$

- Funkcija verodostojnosti je proizvod izraza iz Eq. (1) za svako i:

$$\prod_{i=1}^{n} P(X_i = x_i|\mu). \tag{2}$$

- Da bismo pronašli maksimum funkcije verodostojnosti koristimo izvod Eq. 2 po μ i jednakost nuli. Dobijamo formule za tri estimatora:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\widehat{cov(x,y)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Izvor slike: `https://www.youtube.com/watch?v=LmF6fkdqMcs`

Metoda maksimalne verostojnosti za linearni model

- Our model: $E[Y|X] = aX + b$

$$a = \frac{cov(X,Y)}{var(X)} \qquad b = E[Y] - aE[X]$$

- Our MLE estimators:

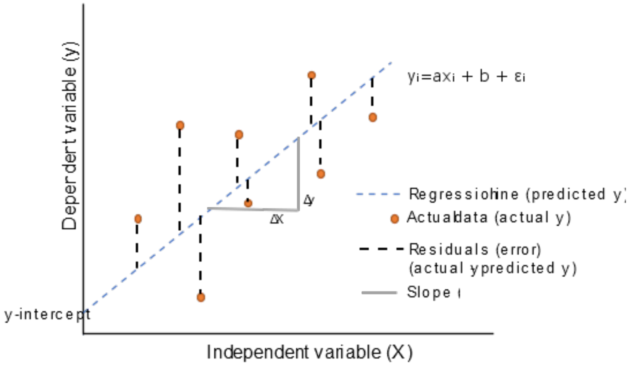
$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

Izvor `https://www.youtube.com/watch?v=LmF6fkdqMcs`

Metoda najmanjih kvadrata

- Linearni model regresije:

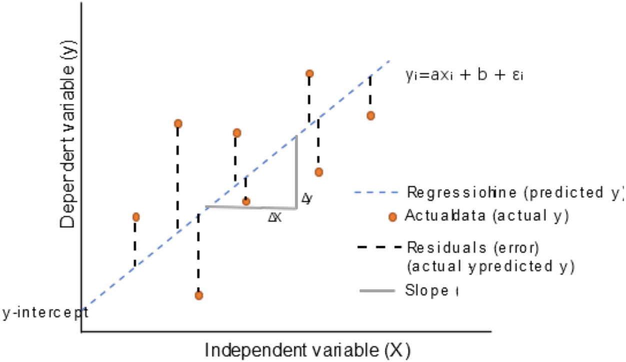
$$y_i = ax_i + b + \epsilon_i$$



Izvor `https://www.reneshbedre.com/blog/linear-regression.html`

- Pretpostavka je da su reziduali nezavisni i $\epsilon_i \sim \mathcal{N}(0, 1)$.

$$y_i = ax_i + b + \epsilon_i$$



Izvor `https://www.reneshbedre.com/blog/linear-regression.html`

Metoda najmanjih kvadrata

- Zadatak je naći koeficijente a i b tako da se minimizira srednjekvadratna greška, odnosno odstupanje stvarnih vrednosti y_i od modeliranih $\hat{y}_i$.
- Srednjekvadratna greška (*eng*. sum of the squares of the residulas (SSE) ili residual sum of squares (RSS)).

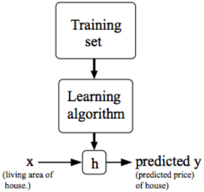
$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

- Ovim dolazimo do problema optimizacije kojeg možemo rešiti na razne načine.
- Rešenja ovog problema minimizacije su estimatori $\hat{a}$ i $\hat{b}$. Podudaraju se sa rešenjima dobijenim metodom maksimalne verodostojnosti.

Terminologija - model|algoritam, atribut|feature

- Iz engleskog nam dolaze dva pojma: *feature* i *atribute*. Postoji razlika.
  - *atribute* - osobina, karakteristika, tip podatka.
  - *feature* - nešto što se gradi (atribut + vrednost).
- Model i algoritam mašinskog učenja. Model je rezultat primene algoritma na ulaznim podacima.

$$Model = Algoritam(podaci).$$



Terminologija - model|algoritam

- Model je rezultat rada algoritma za učenje i predstavlja skup koji sačinjavaju:
  - podaci (brojevi),
  - pravila,
  - strukture podataka specifičine za algoritme.
- Primer.
  - Algoritam linearne regresije proizvodi model koji se sastoji od vektora koeficijenata sa konkretnim vrednostima.
  - Algoritam drva odlučivanja generiše drvo sa if-then uslovima sa konkretnim vrednostima.
  - Algoritam neuralne mreže zajedno sa *backpropagation* i metodom spusta generišu model koji sadrži grafovsku strukturu gde matrice težina imaju specifične vrednosti.
- Na ovaj način imamo:
  MODEL = model data + prediction algorithm.

Linearna regresija u mašinskom učenju - notacija

- Linearni model za jednu promenljivu x: $h(x) = \theta_0 + \theta_1 x$.
- Linearni model za višedimenziona promenljivu $x = (x1, x2)$:

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

- Linearni model za n-dimenzionu promenljivu $\boldsymbol{x}$ (tj. kada imamo n atributa):

$$h(x) = \sum_{i=0}^{n} \theta_i x_i \text{ gde je } x_0 = 1.$$

- Neka je m broj primera u skupu za podučavanje (broj vrsta u tabli, instance).
- x je ulazna promenljiva / svojstvo / atribut.
- y je rezultat / ciljna promenljiva.
- (x, y) je primer za treniranje. $(x^{(i)}, y^{(i)})$ je i-ti primer iz skupa za obučavanje.

Određivanje koeficijenata θ

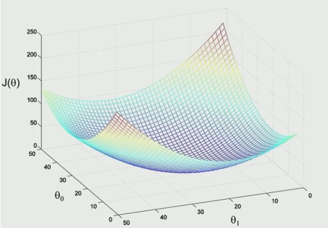
- Izabrati vrednost za θ tako da funkcija $h(x)$ aproksimira vrednosti y,odnosno
$$h(x) \approx y.$$
- Definišemo funkciju greške $J(\theta)$ kao kvadratni kriterijum:
$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2.$$
- Tražimo minimalnu vrednost funkcije cilja (*eng.* cost function) $J(\theta)$:
$$\min_{\theta} J(\theta).$$

Gradijentna metoda, metoda najstrmijeg spusta

1. Pretpostavimo neku inicijalnu vrednost θ_0.
2. Inicijalizujemo parametar λ - *stopa učenja* (*eng*. learning rate).
3. Ažuriramo vrednost za θ_i iterativno formulom:

$$\theta_i := \theta_i - \lambda \sum_{j=1}^{m} (h(x^{(j)}) - y^{(j)}) x_i^{(j)}$$

4. Ponoviti 3. korak do pronalaženja globalnog optimuma.



Izvor slike: `https://www.youtube.com/watch?v=4b4MUYve_U8&t=300s`

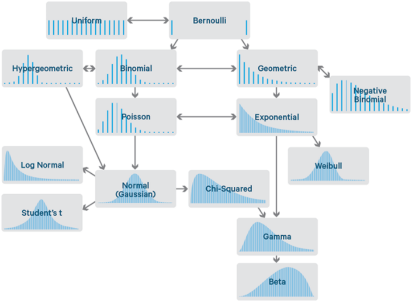
Hvala na pažnji.

Sada idemo na prikaz komandi.

Da li imate pitanja?

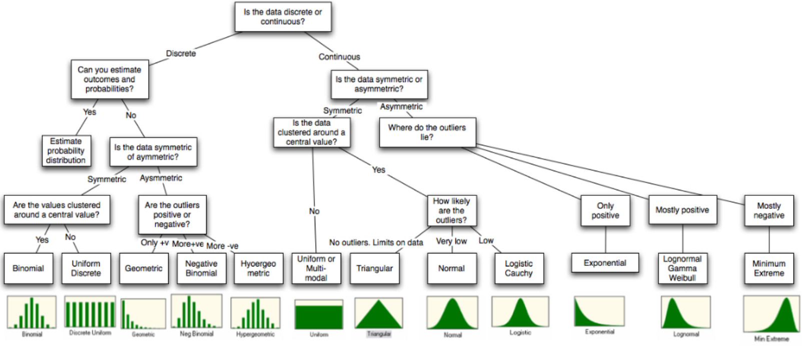
Raspodela podataka

- Prikazujemo raspodelu (distribuciju) podataka u jednom uzorku.
- Histogram.



Izvor `https://medium.com/swlh/understanding-probability-distribution-b5c041f5d564`

Raspodela podataka



Izvor `http://people.stern.nyu.edu/adamodar/New_Home_Page/StatFile/statdistns.htm`